PIM_NER annotálási útmutató

Simon Eszter

2020. október 25.

1. Bevezetés

A számítógépes nyelvészet egy interdiszciplináris terület, amelynek célja az emberi nyelv szerkezetének gépi modellálása, valamint a természetes nyelvek számítógépes feldolgozása. Az információkinyerés a számítógépes nyelvészet egyik fontos alterülete; célja, hogy strukturálatlan szövegből automatikusan hozzájussunk a számunkra értékes információhoz. Mivel ezen információ nagy része tulajdonnevek formájában jelenik meg a szövegben, ezért a tulajdonnév-felismerésnek (named entity recognition, NER) kiemelt jelentősége van.

A NER során egy bemeneti tokensorozatban kell megnevezett entitást (named entity, NE) alkotó intervallumokat kijelölünk, ezeket véges sok kategóriába besorolva. A NE-k nem teljesen azonosak a tulajdonnevekkel, de a jelen annotációs sémában nem lépünk ki a nevek világán kívülre, így a 'NE' és a 'név' szavak szinonimaként fognak szerepelni a tárgyalási univerzumunkban.

A nevek egyedi referenciával bírnak, vagyis a világ egy egyedi entitására utalnak (pl. *London*). Ilyet a köznévi frázisok önmagukban állva (pl. *város*) nem tudnak csinálni, csak akkor, ha egyéb nyelvi elemekkel közösen egy olyan főnévi frázist alkotnak, amely már tényleg a világ egy egyedi entitására utal (pl. *az a város, ahol 9 millió brit lakik*). Mi itt csak a neveket annotáljuk, például:

(1) *Kosztolányi Dezső*

(2) *Szilas Menti Mezőgazdasági Termelőszövetkezet*

(3) *United Nations Educational, Scientific and Cultural Organization*

(4) *Déli-Shetland-szk.*

(5) *IBM*

(6) *Kiss János altábornagy utca*

(7) *Műegyetem*

(8) *The Coca-Cola Co.*

(9) *Kovács Pistike*

2. Az annotálás alapelvei

Fontos, az annotálás során végig szem előtt tartandó szabályok:

- Csak tulajdonneveket annotálunk. Nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem tulajdonnévvel. Például a *József Attila Gimnázium* annotálandó, de a szövegben szereplő az *a suli* frázis nem, hiába derül ki a szövegből, hogy melyik iskolára utal.

- A nevek nem kompozicionálisak. Mivel a nevek jelölete nem a név részeinek a jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy földrajzi névként jelölendő, hiába van benne egy személynév. Ebből az következik, hogy mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.

- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.

- A névannotálásnak két megközelítése lehetséges: a *tag-for-meaning* elv szerint egy nevet mindig az aktuális kontextusnak megfelelő referenciája alapján annotálunk, a *tag-for-tagging* elv szerint pedig mindig az elsődleges referenciája alapján. Ez a két elv egyszerre nem teljesülhet, de mi mégis megkíséreljük összehozni. Alapvetően a *tag-for-tagging* elvet követjük, de külön jelöljük a metonimikusan viselkedő neveket is. A négy alapvető névtípust a 3. fejezetben írjuk le, a metonimikusan viselkedő neveket pedig a 3.5. fejezetben.

- Ha az azonosított név ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.

- A nevek képzett alakjait nem jelöljük. Nem annotálandók tehát az olyanok, mint *magyarországi, fideszes, petőfieskedő.*

- Ha a név összetétel előtagja, és az összetétel alaptagja köznév, például *Horn-kormány, Tilos Rádió-hallgatók, TA-vezérigazgató*, akkor nem annotálandó névként.

- A névhez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, például *The Hague, The Times.*

- A név rövidítése (akronim, mozaikszó) is névként annotálandó.

- A szöveghez a névannotálás során nem nyúlunk hozzá, vagyis nem javítjuk ki a helyesírási hibákat, nem vonunk egybe különírt szavakat, és nem választunk szét egybeírtakat. Ha valamilyen éktelen hibát látunk az aktuálisan címkézendő névvel kapcsolatban, akkor azt külön fel kell jegyezni és elküldeni e-mailben.

3. Alapvető névtípusok

A következő négy alapvető névtípussal dolgozunk:

PERSON: Valós és kitalált személyek neve, becenevek, művésznevek, álnevek.

ORGANIZATION: Olyan csoportok nevei, amelyek valamilyen szervezett struktúrával rendelkeznek, például intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek.

LOCATION: Földrajzilag vagy politikailag definiált helyek nevei, úgymint városok, országok, hegyek, völgyek stb. Idetartoznak az emberalkotta építmények is, mint a repterek, utak, gyárak, épületek stb.

MISC: A felsorolt típusok egyikébe sem tartozó nevek.

Az útmutatóban a NE-ket [szögletes zárójelek] közé tesszük. A példáknál csak az olyan típusú NE-ket jelöljük, amelyikről éppen szó van. A példákban a személyneveket a PER, a szervezetneveket az ORG, a helyneveket a LOC, a be nem sorolhatókat pedig MISC rövidítésekkel jelöljük.

3.1. Személynevek (PERSON)

Személyekre utalhatnak teljes személynevek, becenevek, művésznevek, álnevek, rövidítések. A kitalált személyek, úgymint mozihősök, mesefigurák, mitológiai alakok, illetve a szentek, bibliai alakok nevei is személynévként annotálandók, például:

(10) [Ady_{PER}] *írói álneve* [Ida_{PER}]

(11) *a legkisebb gyerek, aki gyakran játszik* [$Mikrobival_{PER}$]

(12) *zenéjével meglágyította* [$Hádész_{PER}$] *és* [$Perszephoné_{PER}$] *szívét*

A családnevek, az uralkodóházak nevei is személyekre, egészen pontosan személyek csoportjára referálnak, ezért azokat is személynévként kell megjelölni, például:

(13) *a* [$Széchényi_{PER}$] *család Nógrád megyéből származik*

(14) *a* [$Károlyiak_{PER}$] *Apáti nevű faluját felgyújtották*

A személynévhez nem tartoznak hozzá a rangot, címet vagy beosztást jelölő köznévi szavak, például:

(15) *Lemondott* [*Heinz-Christian* $Strache_{PER}$] *osztrák alkancellár*

3.2. Szervezetnevek (ORGANIZATION)

Azok a tulajdonnevek, amelyek egy szervezett struktúrával rendelkező csoportra referálnak, szervezetnévként annotálandók. A következők mind ilyenek:

- Cégek, vállalatok

  (16) *a [SERCO Kft.$_{ORG}$] az eltelt évek során jelentős fejlődésen ment keresztül*

  (17) *1878-ban Grosvernor Lowry-val létrehozzák az [Edison Electric Light Co.-t$_{ORG}$]*

- Multinacionális szervezetek

  (18) *az [Európai Unió$_{ORG}$] ezen a néven 1992-ben jött létre*

- Politikai pártok

  (19) *bántalmazták a [Fidesz$_{ORG}$] egyik ajánlószelvényeket gyűjtő aktivistáját*

- Sportcsapatok

  (20) *A [Budapest Black Knights$_{ORG}$] csapata fölényesen legyőzte a [Szolnok Soldiers$_{ORG}$] csapatát.*

- Katonai szervezetek

  (21) *Az [Észak-atlanti Szerződés Szervezete$_{ORG}$] székhelye Brüsszelben van.*

- Kórházak, egészségügyi intézmények

  (22) *A [Péterfy Kórház-Rendelőintézet Országos Traumatológiai Intézet$_{ORG}$] a főváros egyik legnagyobb egészségügyi intézménye*

- Hotelek

  (23) *A paksiakat is várja az [Erzsébet Nagy Szálloda$_{ORG}$]*

- Színházak, múzeumok

  (24) *A [Szépművészeti Múzeum$_{ORG}$] az egyetemes és a magyar művészet emlékeit mutatja be az ókortól a 18. század végéig.*

- Egyetemek

  (25) *A [Kossuth Lajos Tudományegyetem$_{ORG}$] Honoris Causa Doktorai.*

- Kormányzati hivatalok

  (26) *A [Honvédelmi Minisztérium$_{ORG}$] hivatalos Facebook oldala*

- Szerkesztőségek

  (27) *Mi, akik a [Telexnél$_{ORG}$] dolgozunk, vállaljuk, hogy*

Az általános intézményneveket, mint *rendőrség* vagy *kormány* nem annotáljuk, mert ezek csak egy főnévi frázis részeként tudnak egy egyedi entitást jelölni, nem egyedi jelölők.

Az olyan hosszú, többtagú intézményneveket, amelyek tartalmaznak köznévi elemet is, teljes egészében névnek kell jelölni, például:

(28) *Az [Állami Privatizációs és Vagyonkezelő Rt.$_{ORG}$] zártkörű részvénytársaságként működő állami vállalat volt.*

Ha egy intézménynév után zárójelbe téve szerepel a rövidítése is, akkor a teljes név és a rövidítés külön nevekként kezelendők, a zárójel pedig nem a név része, például:

(29) *Az autóipari óriás [DaimlerChrysler AG$_{ORG}$] ([DC$_{ORG}$]) amerikai részlege*

3.3. Helynevek (LOCATION)

A helynévnek annotálandó entitások közé tartoznak többek között a kontinensek, az országok, a régiók, a városok, a települések, a repterek, az utak, a gyárak, az óceánok, a tengerek, a folyók, a szigetek, a tavak, a nemzeti parkok, a hegyek és a mitikus helyek, például:

(30) [*Franciaországot$_{LOC}$*] *kilenc ország határolja.*

(31) [*Szihalom$_{LOC}$*] *község* [*Heves megye$_{LOC}$*] [*Füzesabonyi kistérségében$_{LOC}$*].

(32) *A* [*Bükki Nemzeti Park$_{LOC}$*] *mintegy 95 százalékát erdő borítja.*

(33) [*Gatwick$_{LOC}$*] *délre,* [*Stansted$_{LOC}$*] *észak-keletre,* [*Luton$_{LOC}$*] *északnyugatra fekszik* [*Londontól$_{LOC}$*].

(34) *Platón dialógusaiban részletesen szól* [*Atlantisz$_{LOC}$*] *szigetéről.*

3.3.1. Összetett kifejezések

Az olyan összetett kifejezésekben, ahol földrajzi nevek vesszővel elválasztva szerepelnek, és a második név nagyobb helyre referál, tehát egyfajta pontosító funkciót tölt be, a neveket külön annotáljuk, például:

(35) [*Los Angeles$_{LOC}$*], [*Calfornia$_{LOC}$*]

(36) [*Budapest$_{LOC}$*], [*Magyarország$_{LOC}$*]

3.3.2. Köznévi tagok

Vannak olyan földrajzi nevek, melyek köznévi utótagot tartalmaznak. A közvetlenül a földrajzi név után álló köznévi frázisok, melyek hivatalosan is a név részei, a névvel együtt annotálandók, mint például az alábbiak:

(37) [*Váci utca*$_{LOC}$]

(38) [*Erzsébet híd*$_{LOC}$]

(39) [*Baranya megye*$_{LOC}$]

(40) [*Duna–Tisza köze*$_{LOC}$]

Nem tartoznak viszont a földrajzi névhez a magyarázó, pontosító funkciójú elemek, illetve az alkalmi jelzők sem, például:

(41) [*Kent*$_{LOC}$] *grófság*

(42) [*New York*$_{LOC}$] *állam*

(43) [*Gyöngyös*$_{LOC}$] *város*

(44) [*Mátra*$_{LOC}$] *hegység*

(45) [*Duna*$_{LOC}$] *folyó*

(46) *az olasz* [*Alpok*$_{LOC}$]

(47) *a lengyel* [*Magas-Tátra*$_{LOC}$]

(48) *a gyönyörű* [*Alpok*$_{LOC}$]

(49) *„Mit nekem te zordon* [*Kárpátoknak*$_{LOC}$]*...”*

3.4. Egyebek (MISC)

Ebbe a kategóriába kerülnek azok, amelyek NE-k, de a felsorolt kategóriák egyikébe sem illenek bele, mint a könyvcímek, újságnevek, konferencianevek, márkanevek, tőzsdeindexek nevei, programozási nyelvek, például:

(50) *A* [*Le Monde*$_{MISC}$] *francia napilap.*

(51) *fedezze fel a* [*Fiat*$_{MISC}$] *modelleket*

(52) *Érdekel a* [*Python*$_{MISC}$] *programozás?*

3.5. Metonimikusan viselkedő nevek

A metonímiákban egy fogalmat vagy dolgot egy másik fogalom vagy dolog jelölésére használunk, például:

(53) *Az embereket sokkolta Vietnam.*

Referenciaátvitel történik: egy névvel az eredeti referens helyett egy másik referensre utalunk. A nevek esetében szabályos referenciaátvitelről van szó, amikor két szemantikai mező között megvan az a megfeleltetés, hogy az egyikbe tartozó név állhat a másik helyett. A metonímiákat az alapján szokták elnevezni, hogy mi áll mi helyett, és hagyományosan kiskapitálissal szedik őket. Így például az 53. példában a *Vietnam* név a PLACE-FOR-EVENT metonímiába sorolható.

Az alábbiakban az eddigi ismertetett négy alapvető névkategória által kifeszített univerzumon belül maradva hozok példákat az egyes metonímiákra. A példákban a metonimikusan viselkedő neveket olyan összetett címkékkel látom el, amelyek első tagja az elsődleges, második tagja a kontextuális referenciát jelöli, és a két tagot egy kettőspont választja el egymástól.

3.5.1. Személynevek

Tipikus, személyneveket érintő metonímia, amikor a művész nevét használjuk egy műve helyett (ARTIST-FOR-ARTFORM), például:

(54) [$Faulknert_{PER:MISC}$] *olvasok.*

3.5.2. Intézménynevek

A szervezetnevek sokszor valójában egy emberi közösségre utalnak. Ilyenkor a szervezetnév aktorként szerepel az adott szövegkontextusban, és olyanokat tud csinálni, mint például döntést hozni, árat emelni, nyilatkozni valamiről (ORGANIZATION-FOR-MEMBERS):

(55) *Az* [$Eötvös\ József\ Gimnázium_{ORG:PER}$] *idén Luxemburgba megy kirándulni.*

(56) *Az* [$IBM_{ORG:PER}$] *ma jelentette be új technológiáját.*

A szervezeteknek sajátjuk, hogy van székhelyük, és gyakran előfordul, hogy a szervezet nevét helymegjelölésként használjuk (ORGANIZATION-FOR-FACILITY), például:

(57) *tűz ütött ki a* [$Kapos\ Hotelben_{ORG:LOC}$]

(58) *elbarikádozták magukat az* [$SZFE\text{-}n_{ORG:LOC}$]

Szintén meglehetősen jellemző, főleg gazdasági rövidhírekben és tőzsdei jelentésekben gyakran előforduló metonímia, amikor a vállalat nevét használjuk a részvényindexük helyett (ORGANIZATION-FOR-INDEX), például:

(59) *A* [$Mol_{ORG:MISC}$] *10 forinttal 6640 forintra, míg a* [$Matáv_{ORG:MISC}$] *1 forinttal 823 forintra csúszott vissza.*

3.5.3. Helynevek

Sokszor a helynevet használjuk, miközben emberek egy csoportjára referálunk (PLACE-FOR-PEOPLE), például:

(60) [$Franciaország_{LOC:PER}$] *korlátozza a politikai menedékjogot.*

Ennek egy tipikus alesete, amikor egy sportcsapatra utalunk egy ország vagy egy város nevével:

(61) [$Olaszország_{LOC:PER}$] *nyerte a foci vébét.*

Amikor egy esemény nagyon erős asszociációs viszonyban áll egy adott hellyel, akkor gyakran a helynévvel utalunk az eseményre (PLACE-FOR-EVENT):

(62) [$Trianon_{LOC:MISC}$] *megítélése a két háború közötti időben*

3.5.4. Egyéb nevek

Erre még nem találtam példát, bár elvileg lehetséges. Ha lesz ilyen, akkor `MISC:PER`, `MISC:LOC` vagy `MISC:ORG` címkével kell jelölni attól függően, hogy milyen típusú név helyett áll a szövegben az a név, amely alapvetően `MISC` címkét kapna.